Title

PhenDisco: phenotype discovery system for the database of genotypes and phenotypes.

Journal

Journal of the American Medical Informatics Association : JAMIA, 21(1)

ISSN

1067-5027

Authors

Doan, Son
Lin, Ko-Wei
Conway, Mike
et al.

Publication Date

2014

DOI

10.1136/amiajnl-2013-001882

Peer reviewed

PhenDisco: phenotype discovery system for the database of genotypes and phenotypes

Son Doan,[1] Ko-Wei Lin,[1] Mike Conway,[1] Lucila Ohno-Machado,[1] Alex Hsieh,[1] Stephanie Feudjio Feupe,[1] Asher Garland,[1] Mindy K Ross,[1] Xiaoqian Jiang,[1] Seena Farzaneh,[1] Rebecca Walker,[1] Neda Alipanah,[1] Jing Zhang,[1] Hua Xu,[2] Hyeon-Eui Kim[1]

▶ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001882).

[1]Division of Biomedical Informatics, University of California San Diego, La Jolla, California, USA
[2]School of Biomedical Informatics, The University of Texas Health Science Center at, Houston, Houston, Texas, USA

Correspondence to
Dr Hyeon-Eui Kim, Division of Biomedical Informatics, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA; hyk038@ucsd.edu

Received 4 April 2013
Revised 7 July 2013
Accepted 29 July 2013
Published Online First 29 August 2013

To cite: Doan S, Lin K-W, Conway M, *et al*. *J Am Med Inform Assoc* 2014;**21**: 31–36.

ABSTRACT

The database of genotypes and phenotypes (dbGaP) developed by the National Center for Biotechnology Information (NCBI) is a resource that contains information on various genome-wide association studies (GWAS) and is currently available via NCBI's dbGaP Entrez interface. The database is an important resource, providing GWAS data that can be used for new exploratory research or cross-study validation by authorized users. However, finding studies relevant to a particular phenotype of interest is challenging, as phenotype information is presented in a non-standardized way. To address this issue, we developed PhenDisco (phenotype discoverer), a new information retrieval system for dbGaP. PhenDisco consists of two main components: (1) text processing tools that standardize phenotype variables and study metadata, and (2) information retrieval tools that support queries from users and return ranked results. In a preliminary comparison involving 18 search scenarios, PhenDisco showed promising performance for both unranked and ranked search comparisons with dbGaP's search engine Entrez. The system can be accessed at http://pfindr.net.

INTRODUCTION

The database of genotypes and phenotypes (dbGaP) is an important repository for data generated through various genome-wide association studies (GWAS), which can be used for new explorations or cross-study validation.[1–3] In addition to genomic data, dbGaP requires investigators to submit phenotype data. As of 7 July 2013, dbGaP contained 422 studies, including more than 130 000 phenotype variables. However, searching relevant studies accurately and completely is challenging, because phenotypic information related to studies is often stored in a non-standardized way. For particular queries, the dbGaP Entrez system returns several studies that are not always relevant, and it does not make clear how particular records are selected and why they appear in a particular order. Consequently, users have to review each study description carefully to determine relevancy, which can become a laborious and time-consuming task when many studies are retrieved.

To address this issue, we developed a new web-based information retrieval system called PhenDisco (phenotype discoverer) based on the user requirements obtained by interviewing dbGaP users. The project is funded through the program entitled phenotype finder in data resources (pFINDR) from the National Heart, Lung, and Blood Institute. The goal of this program is to facilitate the search of phenotypes in dbGaP's GWAS. Our approach uses natural language processing (NLP) as well as information retrieval techniques in order to improve phenotype search in dbGaP.

There are several related works that aim to address issues associated with the lack of standardization in phenotype variables.[3–9] PhenX defined 287 frequently used phenotypes (called measures) in 21 research domains, and manually cross-mapped these measures to phenotype variables in 16 dbGaP studies.[3 4] The goal is to use these measures prospectively, so new studies are described in a standardized way. Another project, eMERGE, used a semi-automated process: users manually search for phenotype variables for specific domains (eg, Alzheimer's disease), and these variables are automatically mapped to standardized vocabularies through a tool called eleMAP. eleMAP outputs are then further curated by users before results can be interpreted.[8 9] Our group was involved in similar work that annotated phenotypes in the gene expression omnibus (GEO),[10] a public gene expression data repository. Human annotators reviewed the papers published using the data available in GEO, then manually identified the phenotype variables and mapped them to the National Cancer Institute thesaurus.[5–7] Although the results of such manual or semi-automated mapping processes tend to be reliable and accurate for small data, the technique is not scalable. Therefore, we developed an algorithmic approach to process the large amount of phenotype variables in dbGaP for standardization.

METHODS

PhenDisco consists of two main components: (1) text processing tools that standardize both phenotype variables and study metadata, and (2) information retrieval tools that support queries from users and return ranked results. Below we describe each component.

Data collection and standardization

We collected information about the GWAS and their phenotype variables from two publicly available dbGaP sources: (1) dbGaP web pages (http://www.ncbi.nlm.nih.gov/gap), and (2) the dbGaP FTP site (ftp://ftp.ncbi.nlm.nih.gov/dbgap). The dbGaP web pages contain information about individual study levels such as study ID, title, description, platforms, and the dbGaP FTP site contains phenotypic

information such as phenotype ID, phenotype description and associated statistics. We developed a crawler to download both types of data. We analyzed 422 studies, which contained 130 000 variables.

Study-level metadata generation

Given that the number of new studies being added every month is small, we focused on automating the standardization of variables, while the abstraction of study data itself was only partially automated. Portions of the study-level metadata are well structured and amenable to full automatic parsing. Study ID, title, number of participants, and study design are automatically extractable study data. We extracted, through manual review, study data such as topic diseases, consent type, institutional review board status, and study locations.[11 12] To standardize the study information, the topic diseases were mapped to the unified medical language system (UMLS)'s concept unique identifiers.[13] We adopted UMLS as a controlled vocabulary in this project based on its comprehensive domain coverage and widespread use in biomedical NLP systems.[14 15] In addition, we mapped study locations to ISO 3166-2 country subdivision code,[16] for example, US-AZ (USA—Arizona).

Phenotype variable standardization

The task of phenotype variable standardization has been the most interesting, yet most challenging, part of developing PhenDisco. The lack of a uniform naming convention meant that, for a study containing thousands of phenotype variables, idiosyncratic choices introduced unnecessary variation and redundancy across studies. For example, the same variable 'body weight' can be represented as 'weight' (variable id: phv00173256.v1.p1), 'WGHT' (variable id: phv00169068.v2.p1), and 'FB9' (variable id: phsv00001189.v1.p7). Therefore, variable descriptions, which provide more information than variable names, are more useful for the task of standardization. The lack of standardization is a well-known problem in clinical informatics; standards and information models, such as the clinical elements model (CEM), were designed to address this issue. The CEM worked reasonably well for clinical variables in electronic medical records, but did not address clinical research variables in dbGaP.[17] While standards such as the observational medical outcome partnership (OMOP) model[18 19] cover many of these variables, given our experience mapping variables into OMOP for a very limited set of conditions,[20] we realized that the variables in dbGaP studies were described in less detail and determined that it would be more cost-effective and scalable to map them into a simpler model.[21 22] We briefly describe our approach as follows.

We developed an information model including four major information classes: 'theme' (ie, age, gender, race, ethnicity), 'subject', 'event', and 'linkage' of information.[21 23 24] For example, the phenotype variable 'age Mom diagnosed—asthma' has theme age, subject 'mother', event 'asthma', and linkage of information 'diagnosed'. We wrote a simple NLP tool in Python called DIVER to identify and map phenotype variables into this model. The evaluation on 3565 variables from pulmonary studies in dbGaP showed that DIVER achieved 98% recall and 94% precision in identifying variables related to demographic concepts and 79% correct mapping into the information model.[23]

For variables that were not related to demographic concepts, we identified two categories of variables: 'topic' and 'subject of information'. The 'topic' is the main theme of phenotype variables while the 'subject of information' is the individual experiencing the variable. For example, the phenotype variable 'father diagnosed with lung cancer' has subject of information 'father' and topic 'lung cancer'. We first tagged 'topic' and 'subject of information' terms from each variable description, and then mapped those terms to the UMLS metathesaurus.[13] This process was automatically implemented by our customized NLP tool. Further standardization of these variables based on information modeling and NLP is in progress.[21]

Information retrieval and ranking algorithm

The information retrieval tool consists of two parts: a query parser and a ranking algorithm.

Query parser

We utilized pyparsing[25]—a toolkit written in Python—for parsing queries in PhenDisco. The role of a query parser is to take an input query and break it into its respective terms and operators. Search terms can be a single word or whole phrases, connected by operators (ie, AND, OR, NOT). To improve search performance, we expanded each input query to include synonyms by integrating MetaMap[26] into the query parser. This concept-based search is the default search mode of PhenDisco (see figure 1).

Ranking algorithm

We used the BM25F ranking algorithm,[27 28] as it is one of the most popular ranking algorithms for structured documents. BM25F is a modified tf-idf (term frequency—inverse document frequency) algorithm[29] that has been shown to enhance performance when dealing with documents composed of several fields such as title, headline, main text.[30 31] We considered each study using the different fields identified in the study abstraction process, such as title, study description, or topic disease, along with standardized phenotypes. In this first version of PhenDisco, we considered terms from different fields to be equally important, and we plan to analyze user searches and rankings to assign appropriate weights for these terms in the next version of the software. We utilized Whoosh,[32] a search library, to implement the BM25F algorithm. The system components are depicted in figure 2. The system is implemented in Linux Ubuntu OS 64-bit using 32GB RAM, running MySQL V.14.14 on an Apache V.2.2.20 web server and is available at http://pfindr.net.

Key system features

Currently, PhenDisco supports basic keyword searches and offers the following features that are not supported in dbGap Entrez:

- Auto-complete: auto-completion of search term function was integrated with the search box, using the phenotype terms collected from the GWAS catalog.[33]
- Concept-based search: search term expansion by synonym based on UMLS metathesaurus mapping.
- Highlighted search keywords: the terms relevant to the search keywords are highlighted in the search result display.
- Ranked results: returned studies are displayed in ranked order, determined by the BM25F algorithm.
- Customization of the result display: users can select the study level metadata such as title, study type, platform to display with the search results. Users can select and export results to the comma-separated values format.

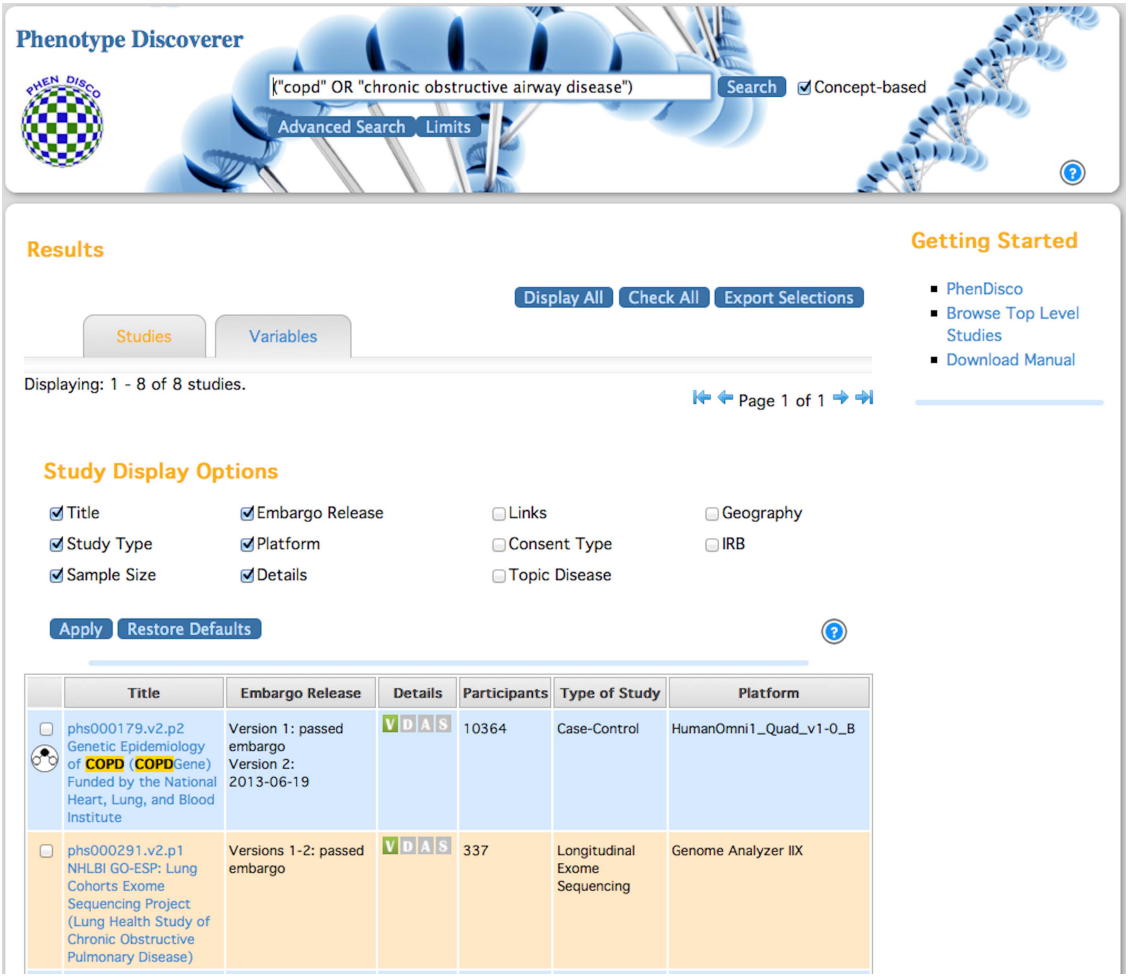

Figure 1 Screenshot of the PhenDisco system. The top panel contains a search input box with concept-based search (ie, expandable terms) as the default.

EVALUATION

Gold standard dataset

A domain expert developed 18 search scenarios related to particular cardiopulmonary conditions. Search scenarios could included disease names such as 'asthma', 'myocardial infarction' in combination with demographics such as 'African American' and/or a clinical attribute such as 'FVC' (forced vital capacity). The list of queries used for evaluation is listed in table 1. Use

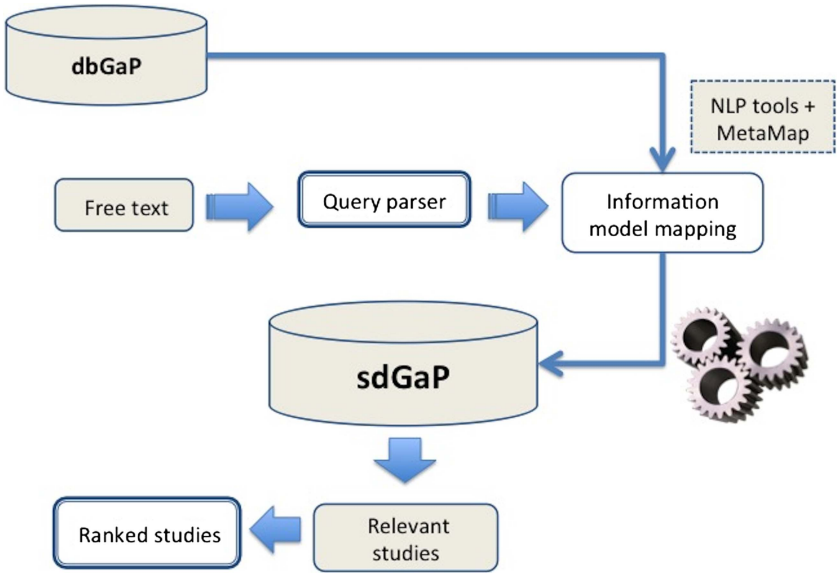

Figure 2 Components of the PhenDisco system: (1) sdGaP (semantic-driven genotypes and phenotype) database contains standardized phenotype variables and study metadata from dbGaP, and (2) information retrieval tools that parse input queries, map into information model and return ranked studies. sdGaP consists of data from dbGaP that are mapped into our information model, as well as study meta-data.

Table 1 List of 18 user-defined queries used for pilot evaluation

Case no.	Query
1	Asthma
2	Asthma AND 'African American'
3	Asthma AND 'African American' AND Hispanic
4	Asthma AND 'African American' AND 'skin test'
5	Asthma AND 'African American' AND Hispanic AND 'skin test'
6	Asthma AND 'African American' AND FEV1
7	Asthma AND 'African American' AND Hispanic AND FEV1
8	Asthma AND 'skin test'
9	COPD
10	'Chronic obstructive pulmonary disease' AND Caucasian
11	'Chronic obstructive pulmonary disease' AND Caucasian AND 'high cholesterol'
12	COPD AND hypercholesterolemia
13	COPD AND FVC
14	'Chronic obstructive pulmonary disease' AND Caucasian AND FVC
15	'Myocardial infarction'
16	'Myocardial infarction' AND black
17	MI AND BMI
18	'Myocardial infarction' AND black AND BMI

cases were determined based on presumed clinical relevance, clinical interest, and potential future research impact. For example, in regard to use cases 1–9, 'asthma' was chosen because of its widespread prevalence.[34]

The domain expert then manually reviewed all dbGaP studies and created the gold standard for each search scenario according to the following steps:

1. Review entire set of dbGaP studies and find studies that were relevant to 'disease' keywords (eg, 'asthma'),
2. review all information resources (ie, study description, phenotype variable description) related to the retrieved studies, and
3. find studies that include demographic information (eg, 'African American') and a clinical attribute (eg, FVC).

Evaluation metrics

We conducted a preliminary evaluation of the system using standard information retrieval measurements: precision, recall and F-measure for unranked studies.[35–37] For relevancy ranking, we used two measures: mean rank precision (MRP) and mean average precision (MAP). They are widely used in information retrieval evaluation for both general and biomedical texts.[38–41] MRP is the mean value of the precisions computed over all queries at a certain cut-off rank. MAP is the mean value of the average precisions for each rank computed for all queries. Average precision is calculated as follows:

$$\text{Average precision} = \frac{\sum_{i=1}^{n} (\text{precision}(i) \times \text{rel}(i))}{\text{number of relevant studies}}$$

Here n is the number of returned documents; precision(i) is the precision at rank i, and rel(i) is an indicator function at rank i: it equals 1 if the corresponding study is relevant, and 0 otherwise. In our evaluation we chose the cut-off rank to be 5, which is a frequently selected cut-off point.[30 38–40]

Table 2 Information retrieval performance of PhenDisco versus dbGaP on 18 user case queries

	Precision	Recall	F-measure	MRP (top 5)	MAP
dbGaP Entrez	0.0756	0.5278	0.1321	0.0600	0.0756
PhenDisco	0.3000	0.9722	0.4552	0.4000	0.2971

MRP (top 5) is mean rank precision at top five retrieved studies, MAP is mean average precision.

PhenDisco performance

Our evaluation of PhenDisco and dbGaP Entrez was conducted on 10 January 2013. The results are shown in table 2 (see more details in supplementary appendix 2, available online only). For the limited number of queries that were evaluated, PhenDisco had substantially better performance than dbGaP Entrez, with an F-measure of 0.4552 versus 0.1321 for the unranked evaluation. When ranking was considered for the top five returns, PhenDisco also showed better performance than dbGaP Entrez with the MRP of 0.40 versus 0.06, and MAP of 0.2971 versus 0.0756.

A preliminary evaluation of usability from three real dbGaP users indicated that PhenDisco fully satisfied the usability requirements they put forward (see more details in supplementary appendix 3, available online only).

DISCUSSION

PhenDisco achieved higher recall and precision than dbGaP in both unranked and ranked results in this pilot evaluation. Through error analysis, we found that dbGaP's low precision was mainly due to its acceptance of search terms that appear in any text in any part of the study, including less relevant contexts such as exclusion criteria or title of papers referenced on the study description. On the other hand, the main reason for the low recall of dbGaP Entrez is the lack of standardization of phenotype information. In other words, dbGaP Entrez only supported string-based search, thus search terms such as 'myocardial infarction' were not expanded into synonymous or acronyms such as 'heart attack' and 'MI'. The fact that dbGaP Entrez returns unranked results accounts for that system's low performance in the relevance ranking evaluation.

Precision in PhenDisco was higher than in dbGaP Entrez, but was still lower than expected. This may have resulted from the utilization of too stringent a criterion to consider a particular study as being 'relevant' for the search. The domain expert was focused on the primary goals of the studies for this formative evaluation, and not on the availability of the phenotype in general (eg, if 'asthma' was not a main subject for a study, then the domain expert considered the study not to be relevant, although the study might have contained individuals with that phenotype and hence it would not be necessarily a false positive). In the comparison between Entrez and PhenDisco, however, using a stringent criterion affected both systems equally. In future work we will investigate the appropriateness of using a less stringent criterion to categorize studies into relevant or not relevant for a particular search. We believe that the best way to categorize may be to obtain direct feedback from users. For example, by unselecting studies that appear in the output, users are indicating that they are irrelevant for their searches. Once we collect data from a large number of users, we will be able to enhance our system and provide more accurate precision and recall estimates.

PhenDisco may be a good alternative to dbGaP Entrez for scientists who need to identify studies that contain the phenotypes they are interested in. Some advantages of PhenDisco over dbGaP Entrez are: (1) PhenDisco integrates NLP tools to enhance query

processing and phenotype variable mapping; (2) PhenDisco augments background knowledge from domain experts by adding meta-data for the studies; and (3) PhenDisco's results are ranked in descending order of relevance. The main disadvantage of PhenDisco is that, unlike dbGaP Entrez, which relies on keyword search in any portion of a study document, PhenDisco's search is performed on study and variable descriptions only, based on meta-data that are produced by a process that is not fully automated. We use a curator to verify a large portion of the results of an automated mapping process and to fix annotations as needed. Given our simple information model, it takes less than 30 min for a curator to validate the majority of the meta-data and this is why we were able to annotate all studies in dbGaP with the help of part-time curators. As the number of new studies is relatively small when compared to over 400 that underwent this process, the semi-automated process is scalable and is not a bottleneck. We plan to improve further the information model and mapping algorithm and use the same process to annotate phenotypes in GEO and other public data resources.

In the future, we plan to add more features to the current system and keep our users updated by prominently displaying the changes in the home page of PhenDisco's web site. These features include: (1) improving the search performance, especially by integrating search queries with ontology expansions for concepts' children; (2) improving PhenDisco's advanced search, by incorporating other types of study level meta-data; (3) providing efficient ways of identifying and browsing similar phenotype variables collected across different studies using clustering techniques. We also plan to apply more sophisticated NLP techniques to improve precision of the system to account for detection of negated concepts and temporal relationships, and promote broader dissemination of the tool and meta-data through the iDASH National Center for Biomedical Computing.[42]

Correction notice This article has been corrected since it was published Online First. The last author's name was previously incorrect and has now been corrected.

Acknowledgements The authors would like to thank Wendy Chapman, Melissa Tharp, Jihoon Kim, and the dbGaP helpdesk for their valuable help and feedback in the early phases of this project. They thank Karen Truong, Myoung Lah, Vinay Venkatesh, Rafael Talavera, and other internship students for their contributions to the system. The authors also thank NIH officers and their scientific advisory board for helpful feedback.

Contributors SD was the main software developer, creating the framework and backend pipeline of the system. He wrote the manuscript with the help of others. H-EK was the main investigator for this work and led phenotype standardization and user requirement analysis. She also contributed to system evaluation. K-WL contributed to phenotype standardization, system evaluation, and user requirement analysis. MC contributed to the study abstraction work and also participated in phenotype standardization development. AG contributed to user interface design and development. SFF contributed to study abstraction, system evaluation, and user requirement analysis. AH mainly contributed to phenotype standardization and user interface development. MKR contributed to study abstraction, ranking algorithm development and system evaluation. XJ contributed to the development of the ranking algorithm. NA contributed to system evaluation and phenotype standardization. HX contributed to phenotype standardization. RW contributed to phenotype standardization, system evaluation and user requirement analysis. SF contributed to phenotype standardization. JZ contributed to system evaluation. LO-M provided oversight for this work and substantial input to this manuscript.

Funding This work was supported in part by grants UH2HL108785, U54HL108460, and T15LM011271 from the NIH.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Mailman MD, Feolo M, Jin Y, *et al*. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet* 2007;39:1181–6.
2 Walker L, Starks H, West KM, *et al*. dbGaP Data Access Requests: A Call for Greater Transparency. *Sci Transl Med* 2011;3:113cm34.
3 Pan H, Tryka KA, Vreeman DJ, *et al*. Using PhenX measures to identify opportunities for cross-study analysis. *Hum Mutat* 2012;33:849–57.
4 Stover PJ, Harlan WR, Hammond JA, *et al*. PhenX: a toolkit for interdisciplinary genetics research. *Curr Opin Lipidol* 2010;21:136–40.
5 Lacson R, Pitzer E, Kim J, *et al*. DSGeo: software tools for cross-platform analysis of gene expression data in GEO. *J Biomed Inform* 2010;43:709–15.
6 Pitzer E, Lacson R, Hinske C, *et al*. Towards large-scale sample annotation in gene expression repositories. *BMC Bioinformatics* 2009;10:S9.
7 Lacson R, Pitzer E, Hinske C, *et al*. Evaluation of a large-scale biomedical data annotation initiative. *BMC Bioinformatics* 2009;10:S10.
8 Pathak J, Pan H, Wang J, *et al*. Evaluating Phenotypic Data Elements for Genetics and Epidemiological Research: Experiences from the eMERGE and PhenX Network Projects. In: *AMIA Summits on Translational Science proceedings AMIA Summit on Translational Science*. 2011:41–5.
9 Pathak J, Wang J, Kashyap S, *et al*. Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. *J Am Med Inform Assoc* 2011;18:376–86.
10 Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 2002;30:207–10.
11 Truong K, Conway M. A Study on Studies: Exploring the Metadata Associated with dbGaP Studies. In: *IEEE Second Conference on Healthcare Informatics, Imaging and Systems Biology*. 2012:126.
12 Ross MK, Lin KW, Truong K, *et al*. Text Categorization of Heart, Lung, and Blood Studies in the Database of Genotypes and Phenotypes (dbGaP) Utilizing n-grams and Metadata Features. *Biomed Inform Insights* 2013;6:35–45.
13 Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Methods Inf Med* 1993;32:281–91.
14 Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004;32:D267–70.
15 Kleinsorge R, Tilley C, Willis J. Unified Medical Language System (UMLS). In: Kent A, Hall CM, eds. *Encyclopedia of Library and Information Science*. Marcel Dekker, 2002:369–78.
16 (ISO) International Organization for Standardization. ISO 3166-2. http://www.iso.org/iso/home/standards/country_codes.htm.
17 Lin K-W, Tharp M, Conway M, *et al*. Feasibility of Using Clinical Element Models (CEM) to Standardize Phenotype Variables in the Database of Genotypes and Phenotypes (dbGaP). In: *IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*. 2012:123.
18 Ryan PB, Madigan D, Stang PE, *et al*. Empirical assessment of methods for risk identification in healthcare data: results from the experiments of the Observational Medical Outcomes Partnership. *Stat Med* 2012;31:4401–15.
19 Li X, Hui S, Ryan P, *et al*. Statistical visualization for assessing performance of methods for safety surveillance using electronic databases. *Pharmacoepidemiol Drug Saf*. Published Online First: 14 February 2013. doi:10.1002/pds.3419
20 Ogunyemi OI, Meeker D, Kim H-E, *et al*. Identifying appropriate reference data models for comparative effectiveness research (CER) studies based on data from clinical information systems. *Med Care*. Published Online First: 13 June 2013. doi:10.1097/MLR.0b013e31829b1e0b
21 Lin K-W, Hsieh A, Farzaneh S, *et al*. Standardizing Phenotype Variables in the Database of Genotypes and phenotypes (dbGaP) based on Information Models. In: *AMIA Summit on Translational Bioinformatics*. 2013:110.
22 Alipanah N, Lin K, Venkatesh V, *et al*. Phenotype Information Retrieval for Existing GWAS Studies. In: *AMIA Summit on Clinical Research Informatics*. 2013:4–8.
23 Hsieh A, Doan S, Conway M, *et al*. Demographics Identification: Variable Extraction Resource (DIVER). In: *2012 IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology*. Ieee, 2012:40–9.
24 Hsieh A, Conway M, Kim H. Identifying Age Variables in dbGaP using Natural Language Processing. In: *AMIA Annu Symp Proc*. 2012:1781.
25 Pyparsing toolkit. http://sourceforge.net/projects/pyparsing/.
26 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In: *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001:17–21.
27 Robertson SE, Walker S, Jones S, *et al*. Okapi at TREC-3. In: *Proceedings of the Third Text REtrieval Conference (TREC 1994)*. Gaithersburg, USA, 1994.
28 Robertson SE, Walker S, Hancock-Beaulieu M. Okapi at TREC-7. In: *Proceedings of the Seventh Text REtrieval Conference*. 1998.
29 Jones KS, Walker S, Robertson SE. A probabilistic model of information retrieval: development and comparative experiments (parts 1 and 2). *Info Process Manag* 2000;36:779–840.

30 Pérez-Agüera JR, Arroyo J, Greenberg J, *et al*. Using BM25F for semantic search. In: *Proceedings of the 3rd International Semantic Search Workshop*. ACM, 2010: 1–8.
31 Robertson S. The probabilistic relevance framework: BM25 and beyond. *Foundations Trends Info Retrieval* 2010;3:333–89.
32 Chaput M. Whoosh 2.4.1. http://pypi.python.org/pypi/Whoosh/.
33 Hindorff L, MacArthur J, Morales J, *et al*. A Catalog of Published Genome-Wide Association Studies. http://www.genome.gov/gwastudies.
34 Centers for Disease Control and Prevention (CDC). Asthma. 2012. http://www.cdc.gov/asthma/.
35 Wikipedia. Precision and recall. http://en.wikipedia.org/wiki/Precision_and_recall (accessed 10 Mar 2013).
36 Van Rijsbergen CJ. *Information Retrieval*. 2nd edn. Butterworth-Heinemann, 1979.
37 Manning CD, Raghavan P, Schütze H. *Introduction to Information Retrieval*. Cambridge University Press, 2008. doi:10.1109/LPT.2009.2020494
38 Lu Z, Kim W, Wilbur WJ. Evaluating relevance ranking strategies for MEDLINE retrieval. *J Am Med Inform Assoc* 2009;16:32–6.
39 Lu Z, Kim W, Wilbur WJ. Evaluation of query expansion using MeSH in PubMed. *J Am Med Inform Assoc* 2009;12:69–80.
40 Hersh W. Evaluation of biomedical text-mining systems: lessons learned from information retrieval. *Brief Bioinform* 2005;6:344–56.
41 Hersh W. *Information retrieval: a health and biomedical perspective*. Springer, 2008.
42 Ohno-Machado L, Bafna C, Boxwala A, *et al*. iDASH. Integrating data for analysis, anonymization, and sharing. *J Am Med Inf Assoc* 2012;19:196–201.